DOCUMENT RESUME

ED 420 683	TM 028 358
AUTHOR	Hearne, Jill; Ramey, Madelaine
TITLE	Addressing Validity Issues in Student Exit Performance Assessment.
PUB DATE	1998-04-00
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; Educational Change; Educational Policy; *Elementary School Students; *Equal Education; Grade 4; *High Risk Students; Intermediate Grades; Low Achievement; Minority Groups; *Performance Based Assessment; Professional Development; *Selection; Sex Differences; Socioeconomic Status; Standards; State Programs; Testing Programs; Urban Schools; *Validity
IDENTIFIERS	*Exit Examinations

ABSTRACT

A large urban northwest school district is engaged in the construction of a central accountability policy that depends on school-based implementation. As part of this effort, the district has implemented a standards-based exit policy. Each school has identified students as either meeting grade-level standards or not meeting these standards. This paper describes an investigation of the application of the policy at various schools and its impact on the students identified as not having the skills to exit a grade. The district has identified 534 fourth graders without the skills to exit. Issues of disproportionately low achievement among minority groups, particularly African Americans, are of central concern in this district. A study examined: (1) the validity of teacher judgments; (2) the relationship between teacher judgments based on classroom evidence and district and state evidence; (3) error rate disproportionalities; and (4) error rates by school. The validity of teacher judgment was influenced by differences between schools in the rates of implementation and use of classroom evidence, although professional training in assessment was creating some uniformity in classroom evidence. The relationships between classroom assessments and state assessments indicate that teachers may rely on demonstrated computational skills as evidence of student achievement, based on the use of the mandated state test. The most striking finding was the disproportionality of error rate by ethnic group, gender, and socioeconomic status. Implications of these findings for educational equity in this urban district are discussed. (Contains 29 references.) (SLD)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

ED 420 683

Addressing Validity Issues In Student Exit Performance Assessment

Jill Hearne
Seattle Public Schools
815 Fourth Avenue North
Seattle, WA 98109
Phone: 206-298-7235
email: jhearne@cks.ssd.k12.wa.us

and

Madelaine Ramey
Rain City Associates
13545 Corliss Avenue North
Seattle, WA 98133
Phone: 206-364-0262

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Jill Hearne

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Presented at the

American Educational Research Association

San Diego, California

April 1998

TM028358

Addressing Validity Issues in
Student Exit Performance Assessment

Jill Hearne, Seattle
(WA) Public Schools;
Madelaine Ramey,
Rain City Associates

Objectives

Issues of standard setting lie at the intersection of statistical reliability and validity and politics and policy making. Standard setting is not an absolute procedure, but one that has been influenced by history (Madaus, 1992) and politics (Linn, 1994). What standards (or standards of performance) define the necessary set of skills to be successful at the next level of schooling and/or successful as a self-determined citizen in a democracy? What is the similarity (or difference) between these standards and professional and public perceptions of competence? And finally, what is the relationship of teacher judgment of standards attainment and more conventional norm-referenced measures?

These are the questions and issues which surround the design of a high stakes exit policy as one district constructs a model of identification and assistance for all students in attaining standards of proficiency in reading, mathematics, and language arts. In a large urban northwest district, a cohort of students in the fourth grade were studied to determine the extent to which their teachers' judgments of proficiency related to multiple measures of student achievement, including district and state norm-referenced tests, teacher judgments based on classroom evidence, and writing tasks. The study examines the following questions:

1. What is the validity of teacher judgments, based on classroom evidence, for decision making in a large urban district's exit profile system?
2. What is the relationship between teacher judgments based on classroom evidence and district and state evidence of student achievement?
3. If a disagreement between classroom evidence and district/state evidence is called "error", do error rates reflect ethnic, gender, and/or socioeconomic class disproportionalities?[1]
4. Do error rate disproportionalities vary by school?

Perspective

After several turbulent years engaged in issues of decentralization and restructuring, a large urban district is engaged in construction of a centralized accountability policy that depends on school-based implementation. This top-down, bottom-up approach to policy implementation is suggested by Elmore and Associates (1991) as necessary in systems with little centralized power and control over constituents.

The advent of a strong superintendent and a cohesive board made it possible to put into place a policy which holds individual schools responsible for individual student achievement with

[1] According to district usage, disproportionality on a given measure occurs when certain ethnic, cultural, or socioeconomic groups are over- or under-represented in comparison with their representation in the total district population.

standards of performance being demonstrated through classroom evidence and supplemented by state and district level assessments.

Inherent in any policy construction surrounding higher standards are the issues of reliability and validity of teacher judgment. Are teacher judgments about student proficiency fair and equitable to all students or are they tainted by "limiting beliefs about differential ability to learn and self-defeating teaching methods that follow from such beliefs?" (Weinstein, 1996). If there is a differentiated standard that shifts in response to perceived student potential (Weinstein, 1996), then grade advancement decisions that rest on classroom based assessments might increase the likelihood that marginalized groups are overrepresented in the population not advanced. Student achievement can be affected by teacher beliefs when teachers reduce the amount and level of schoolwork given students when this reduction is not necessary (Goldenberg and Gallenmore, 1991).

The 1997-1998 school year is the first year of implementing this standards-based exit policy, and each school has identified students as either: (a) yes, meets grade level standards or (b) no, does not meet grade level standards. There are 534 students in 4th grade who have been identified as not on track to exit 5th grade. This paper describes an investigation of the application of the policy at various schools and its impact on the students identified as not having the skills to exit. Issues of disproportionately low achievement among minority groups, particularly African American males, are of central concern in this district.

Methods

Each of the following subsections describes the method used to answer, in order, one of the four questions posed at the beginning of this report. The subsections are headed according to the content of the question: validity of classroom evidence, the relationship between classroom evidence and district/state evidence, error rate disproportionalities, and error rates by school.

Validity of Teacher Judgments

The district is currently engaged in intensive efforts to train teachers in gathering evidence that will be valid for making decisions, particularly those decisions relating to student exit. The steps taken to assure that the evidence is in fact valid were documented by an outside consultant, published within the district, and summarized in the results section of this paper.

Relationship between Teacher Judgments and District/State Evidence

The relationship between classroom evidence and district/state evidence was determined by addressing the following three questions.

- What are the correlations between the District and State measures within each subject area, reading, math, and writing?
- How validly do the District and State measures predict teacher judgments that students meet exit standards based on classroom evidence?
- If the disagreement between teacher judgment that a student meets an exit standard and the student's classification based on a predictor test score is called "error", what is the error rate associated with each test?

The District and State Measures and Their Correlations

The District measures were two subtests of a nationally standardized test, the Iowa Tests of Basic Skills (ITBS) Total Reading and Total Math and a writing performance assessment, referred to as the Direct Writing Assessment (DWA). The ITBS is administered annually, each Spring, to all grades. The DWA is administered in spring to 3rd, 5th, 8th and 11th grade students. The study sample's 3rd grade DWA scores were used in analyses.

The State measures are of recent development, and are being phased in grade by grade. Schools'participation in the 1997 administration was voluntary; about half the schools with a 4th grade participated. The State measures are called the Washington Assessment of Student Learning (WASL), and include reading, math, and writing. The writing test is a performance assessment.

In summary, scores from three District (ITBS Reading, ITBS Math, and DWA) and three State (WASL Reading, WASL Math, and WASL Writing) tests were used in the study. All but one were obtained in spring 1997 while the study sample was in the 4th grade. The one exception (DWA) was from the previous year, 1996.

Validity of District and State Measures for Predicting Teacher Judgments that Students Meet Exit Standards Based on Classroom Evidence

The validity study used discriminant analyses.

Predictor variables. The predictor variables were scores on the District and State test. Normal Curve Equivalent scores (NCEs) were used for theITBS. There was a separate discriminant analysis for each predictor.

Criterion or outcome variables. The criterion variables were categorical variables, teacher judgment that a student met the exit standard for reading--for which the two reading tests were predictors, math—for which the two math tests were predictors, and writing—for which the two writing performance assessments were predictors. The two categories of teacher judgment were: "Yes" (1), meets the standard, and "No" (0), does not meet the standard.

Case selection. Half of the students were in the discriminant analysis. The other half were in the cross validation sample.

Discriminant analysis. The analysis was similar to multiple regression analysis. Since the criteria were dichotomous variables scored 1 (Yes) or 0 (No), the measure of relationship, eta, between the predictors and the criterion (the canonical correlation) is equal to R, the multiple correlation between the predictors and criterion in a multiple regression analysis. The coefficients (B) in the linear discriminant function (Equation 1), the major output of discriminant analysis, are proportional to multiple regression coefficients.

$$D = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_p X_p \qquad \text{(Equation 1)}$$

Error Rates

The linear discriminant function was the basis for classifying students. If a student's discriminant score was closer to the mean for the Yes, meets standards, category, he was classified in that category. If his score was closer to the mean for the No, does not meet standards, category, he was classified in the No category.

Since the distributions of discriminant scores for the two categories overlapped, there were classification errors. There are two kinds of errors: classifying a student in the Yes category on the basis of test score when he was already judged No by the teacherand classifying a student in the No category on the basis of test score when he was judged Yes by the teacher. Likewise, there are two kinds of correct classifications.

Since we already knew the category to which a student belongs, we could count the errors of classification as well as the correct classifications. Since the error rates—the proportions of teacher judgments that disagreed with classification by test score-- were nearly identical for the analysis and the cross validation sample, the two samples were combined to calculate the error rates reported here.

Error Rate Disproportionalities

Errors were distinguished as being one of two types. The first type occurred when the teacher judged that a student met the standard but the student 's test score was below the discriminant analysis cut point. The second type occurred when the teacher judged that a student did not meet the standard but the student's test score was above the discriminant analysis cut point.

Student test score files were merged with files containing information on students' ethnic group membership (white versus nonwhite), gender, and socioeconomic status—as indexed by eligibility/non-eligibility for free or reduced price lunch (FRL). This permitted the calculation of error rates, of each type, for each ethnic-gender-FRL group.

Error Rates by School

The merged data also contained student's school so that school was used as a fourth grouping variable (in addition to ethnicity, gender, and free or reduced price lunch eligibility). This permitted the calculation, within school, of error rates for each ethnic-gender-FRL group.

Results

Given below are the results associated with each of the four questions posed at the beginning of this report. The questions concern: validity of teacher judgments, the relationship between classroom evidence and district/state evidence, error rate disproportionalities, and error rates by school.

Validity of Teacher Judgments

Intensive staff development occurred in the 1997-1998 school year to assure that the evidence used by the teachers is valid for the exit standards as outlined in state-defined grade level standards as criteria for training teachers. Teachers received training in how to judge their students' progress in relationship to the expectations of the WASL at grades 4, 7, and 10 with the proficiency levels at those grades being considered "good enough" for grades 5, 8, and 11. Each school is responsible for maintaining a record somewhere about what permissible evidence was used in making an exit decision about a student.

The district is currently training teachers in classroom assessment, using state endorsed and/or developed materials. Teachers in attendance are those appointed by their schools as instructional/curriculum leaders who pass the training on to other teachers in their schools.

There are two parts to the training. The first part draws from Stiggins' Student-Centered Classroom Assessment (1997). The second part is based on the State of Washington's Primary Classroom-Based Assessment Took Kit developed by the Washington Commission on Student Learning (1997). The following is a brief description of the two parts of the training; Stiggins and Tool Kit, as they relate to the validity of assessing students' achievement of exit standards.

Stiggins portion of the training

The Stiggins portion of the training emphasizes "targets". That is, measurement should be aligned with "clearly visioned" targets. To justify a broad statement like "the student can read and comprehend within the grade level" teachers teach to, and assess the entire range of skills, knowledge, etc., that make up reading and comprehending, i.e., that sample the entire domain. The training, following Stiggins, focuses on primary reading fluency, writing, and mathematics.

- In reading, Stiggins cites work of Pinnell, Pikulski, Wixson, Campbell, Gough, and Beatty (1995) commissioned by the NAEP. The work identifies three elements of reading and includes a 4-point reading fluency scale that assesses all three elements.
- In writing, Stiggins cites Spandel's (1994) checklists for each of five developmental stages of writing. The checklists "permit teachers to locate their students in a developmental continuum" (Stiggins, 1997, p. 291).
- In mathematics, Stiggins describes a teachers' classroom assessment guide developed by the Illinois State Board of Education (1995) for Grade 3. Included in the discussion is a guide for scoring student performance. Each of three dimensions, mathematical knowledge, strategic knowledge, and communication, is scored on a 4-point scale, which is reproduced in Stiggins (1997, pp. 296-297).

Tool Kit

Thus, the Stiggins portion of the training covers the validity of classroom assessment and is related to the exit standards. The tool kit portion provides more specificity and includes the following.

- Grade-level frameworks for reading, writing, mathematics.
- Examples of assessment tools.
- Performance tasks, including those written by K-4 teachers in Washington.
- Examples of scoring criteria.
- Examples of student work.
- Instructional models for collecting classroom-based evidence of student work.

Since Seattle's exit standards are aligned with the state's frameworks for the content areas of reading, writing, and math, teacher training addresses the validity of classroom assessment of student achievement of the exit standards

Relationship between Teacher Judgments and District/State Evidence

The relationship between teacher judgments and district/state evidence was studied by answering the three questions stated near the beginning of the Methods section. It should first be noted that the study used teacher judgment that students met the exit standard for reading based on classroom evidence as the criterion for the Reading subtests. Similarly, teacher judgment that students met the exit standard for math based on classroom evidence was the criterion for the Math subtests. Finally, teacher judgment that students met the exit standard for writing based on classroom evidence was the criterion for the Writing subtests.

1. What are the correlations between the District and State measures within each subject area, reading, math, and writing?

The correlation coefficients are: .79 between WASL Reading and ITBS Reading; .81 between WASL Math and ITBS Math; and .50 between WASL Writing and DWA. The first two coefficients are of a respectable size. The last is so low as to suggest imperfection in one or both of the writing tests. It should be noted that the reliability (in a generalizability sense) is .72 for WASL writing and unknown for DWA.

2. How validly can the District and State measures predict teacher judgments that students meet exit standards based on classroom evidence?

The predictive validity of the WASL, ITBS, and DWA can be seen by noting the canonical correlation coefficients in Table 1. The highest correlation was .60 for ITBS Math. The lowest correlations were for the two writing tests. These correlations are moderate with, for example, ITBS Math (correlation of .60) accounting for only 36% of the variance of the criterion (teacher judgment that students met the exit standard for math based on classroom evidence). The DWA accounted for no more then 18% of the variance of teacher judgment that students met the exit standard for writing based on classroom evidence.

Table 1

Canonical Correlations and Percentages Correctly Classified

Test	Canonical r	Classification Error Rate
WASL Reading	.58	.15
ITBS Reading	.58	.15
WASL Math	.55	.19
ITBS Math	.60	.16
WASL Writing	.48	.22
DWA	.43	.23

Thus only a moderate degree of validity for the tests has been established using teacher judgments that students met exit standards based on classroom evidence as criterion measures. The correlations might be low because of imperfect predictors or criteria or both. Also, the reader is reminded that generalizability is relatively low for WASL writing and has not been established for the DWA.

Another way of looking at the relationship between teacher judgment and student performance on a test is in terms of misclassification rate. How often do teacher judgments disagree with students' classification based on the test?

3. If the disagreement between teacher judgment that a student meets an exit standard and the student's classification based on a predictor test score is called "error", what is the error rate associated with each test?

The third column, classification error rate, in Table 1 gives the error rate associated with each test. As can be seen, the lowest error rates are associated with the two reading tests and the highest error rates are associated with the two writing tests, followed by WASL Math.

Error Rate Disproportionalities

Tables 2 and 3 summarize the findings regarding errors of the first and second type by the demographic variables, ethnicity, gender, and free or reduced price lunch eligibility. From Table 2 it can be seen that errors of the first type, where teacher judges that student meets standard but student's test score is below cut point occurs most often for nonwhite students (male and female) who are eligible for free or reduced price lunch. This result is consistent across tests, except those for writing.

Table 3 shows that errors of the second type, where teacher judges that student does not meet standard although student's test score is above cutpoint, occurs most often with nonwhite male students who are eligible for free or reduced price lunch. This result is consistent across tests, except for WASL reading.

Taken together, the results shown in Tables 2 and 3 show that teacher judgments differ from those indicated by a test score classification most often when students are nonwhite males who are eligible for free or reduced price lunch. The results also suggest a similar, but less pronounced tendency, with respect to nonwhite females who are eligible for free or reduced price lunch. These finding suggest that teachers are less able to diagnose the skill level of students in these groups.

Error Rates by School

There are 68 district schools that have a 4th grade. Of these, 19 had significant rates of Type 1 error--teacher judges that student meets standard, but student's test score is below cut point--for nonwhite students who are eligible for free or reduced price lunch (male or female or both) on 2 or more of the 6 tests. Twenty-two had significant rates of Type 1 error--teacher judges that student does not meet standard but student's test score is above cut point--for the same group(s) on 2 or more of the 6 tests. Two schools had significant error rates of both types on two or more of the tests.

Table 2

Rates of Type 1 Error (Teacher Judges That Student Meets Standard but Student's Test Score is Below Cut Point) by Demographic Variables and Test

Group	WASL Reading	ITBS Reading	WASL Math	ITBS Math	WASL Writing	DWA
White Male not FRL-eligible	.03	.02	.06	.04	.03	.12
White Male FRL-eligible	.09	.05	.12	.07	.09	.13
Nonwhite Male not FRL-eligible	.04	.05	.03	.05	.10	.11
Nonwhite Male FRL-eligible	.10 *	.13 *	.15 *	.12 *	.12	.13
White Female not FRL-eligible	.03	.02	.04	.06	.07	.07
White Female FRL-eligible	.07	09	.08	.09	.11	.15
Nonwhite Female not FRL-eligible	09	.10 *	.10	.11 *	.10	.09
Nonwhite Female FRL-eligible	.14 *	.13 *	.18 *	.12 *	.11	.13
Grand Mean	.07	.07	.10	.08	.09	.11

Note. Asterisk (*) indicates that the error rate is significantly greater (p < .05) than the grand mean error rate.

ERIC Full Text Provided by ERIC

Table 3

Rates of Type 2 Error (Teacher Judges That Student Does Not Meet Standard but Student's Test Score is Above Cut Point) by Demographic Variables and Test

Group	WASL Reading	ITBS Reading	WASL Math	ITBS Math	WASL Writing	DWA
White Male not FRL-eligible	.05	.06	.08	.05	.10	.08
White Male FRL-eligible	.09	.14	.07	.12	.19	.20*
Nonwhite Male not FRL-eligible	.09	.09	.08	.06	.11	.09
Nonwhite Male FRL-eligible	.09	.11 *	.12 *	.11 *	.22 *	.19 *
White Female not FRL-eligible	.03	.03	.07	.05	.07	.04
White Female FRL-eligible	.16*	.08	.10	.09	.13	.15
Nonwhite Female not FRL-eligible	.06	.05	.06	.06	.08	.05
Nonwhite Female FRL-eligible	.11 *	.09	.10*	.08	.15	.18
Grand Mean	.08	.07	.09	.07	.13	.12

Note. Asterisk (*) indicates that the error rate is significantly greater ($p < .05$) than the grand mean error rate.

Discussion

The first question related to the validity of teacher judgments based on classroom evidence. The validity of teacher judgments is influenced by differences between schools in rates of implementation and use of classroom evidence. Schools in which teachers utilized scoring guides and rubrics in collecting classroom based evidence of student achievement were more likely to have clearer expectations for learning targets. The use of Stiggins' Student Centered Classroom

Assessment in conjunction with study groups and participation in assessment literacy classes created some uniformity regarding the purposes and uses of classroom based evidence.

Classroom evidence has been the primary source of data used to make advancement decisions at grade 5. Whether or not a student would leave elementary school to go to middle school was a function of teacher judgment, classroom based evidence and a school team decision made after consulting state and district level test information as well.

Classroom based evidence included daily, weekly and unit work reflective of district adopted text as well as teacher made instruments. The State of Washington also provided each elementary school with a "tool kit" which contains tasks that mirror good assessment methods.

By emphasizing the alignment of Seattle's exit standards with the state frameworks for the content areas of reading, writing, and math, teacher training addressed the validity of classroom assessment of student achievement of the standards. With 70 sites operating in a decentralized structure, it will be necessary to continue to monitor teachers' classroom evidence to insure that this training has been incorporated into daily practice. It needs to become a part of teachers' approach to assessing students' achievement of the exit standards.

The second question examined the relationship between teacher judgments based on classroom evidence and District/State evidence using teacher judgment as the criterion. The respectable correlation coefficients between the ITBS and WASL in Reading and Math (.79 and .81) respectively indicate that these external assessments measured the same kind of student achievements. The low correlation (.50) between the DWA and WASL indicate that one or both measures are imperfect and/or they measure different aspects of performance.

The predictive validity coefficients for the various district and state assessments, varied from .60 (ITBS math) to .43 (DWA). Corresponding classification error rates varied from .15 to .23. In particular, as noted in Table 1, the error rate (.19) relative to the WASL math might highlight teacher misperceptions of math achievement as teachers have traditionally focused on computational skill as an indicator of math ability rather that the more robust measures which align with National Council of Teachers of Mathematics (NCTM) standards that are measured by the WASL math. Correspondingly, the higher correlation between ITBS math (.60) and teacher judgment is possibly due to teacher familiarity with the ITBS subtests and their level of comfort with the strong computational orientation of the test.

Indeed, in schools where large numbers of minority students are present, students might be judged proficient by their teachers specifically on the basis of computational skills. This tendency was noted by Koretz, Linn, Dunbar and Shepard (1991). Their analysis showed that some teachers of minority students tend to focus their mathematics and reading curriculum on content specific to the mandated test thereby limiting the range of instruction made available to minority students to a purely immediate functional level. It is likely, therefore, that for many students, this mandated test, ITBS, influenced teachers to focus on computational rather than problem solving activities.

The most striking finding in this study has been the disproportionality of error rate by ethnic group, gender, and socioeconomic status. There were two types of errors, Type 1 and Type 2. Type 1 error occurred when teachers indicated students met the standard and could proceed to the next grade when district and state evidence indicated they did not have skills. Type 2 error refers to instances where district and state evidence suggests students do have skills, but teachers judge that they do not.

The Type 1 error rate varied from a low of .02 for white males not eligible for free or reduced price lunch to a high of .18 for nonwhite females who were eligible for free or reduced price lunch. As Type 1 error rate reflects the likelihood that teachers are passing on students who do not have skill, it seems that teachers are significantly more likely to pass on minority students (male and female) without skills who are eligible for free or reduced price lunch. There seems also a tendency to pass on minority females without skills who are not eligible for free or reduced price lunch.

Early research on teacher student interaction highlighted the phenomenon that in regard to student achievement it is not just the existence of an expectation that causes self-fulfillment, it is the behavior that the expectation produces. Because teachers expect less, students achieve less. (Brophy and Good, 1994).

This indication of low expectations has been well documented in the literature and creates a self-fulfilling prophecy. It is referred to as "Matthew effects", after Matthew 25:29. "Those who have, will get more until they grow rich while those who have not, will lose even the little they have." (Bronfenbrenner, 1988.) Low expectations limit the opportunities for appropriate instructional intervention if a student is advanced to a high grade in school without an academic support structure. Those perceived as good readers, writers and thinkers are provided both increased opportunities to read write and think critically. Those designated as poor readers or disabled learners are given fewer opportunities to read write or think critically because it is assumed that they are not ready to do what the "able learners" are doing. (Bartoli, 1995)

It is possible that Type 1 errors actually result from test bias, that for some students their ability to perform less well in testing situations was due to factors other than their actual skill level. Perhaps these students did perform well in daily classroom work.

Because standardized tests reflect the majority culture, minority student performance on them may not yield a fair representation of what these students really know and can do, given their economic and educational disadvantages. Lomax, West, Harmon, Viator, and Madaus (1995) found that this unfairness a) makes minority students ineligible for courses necessary for high education, and b) tracks them into groups emphasizing basic skill, rote memorization, and the use of test-like problems in class activities and extensive test preparation rather than higher level thinking skills. Therefore, since both Type 1 and Type 2 errors occur, and norm referenced test scores are presumed to under reflect minority achievement; then Type 2 errors are most egregious.

The more troubling error from an ethical point of view, is the Type 2 error rate. The Type 2 error refers to the teacher judgment of students not meeting standards when state and district measures indicate they do have the skill. Type 2 errors varied from a low of .03 for white females not eligible for free/reduced price lunch to a high of .22 for the WASL writing for nonwhite males eligible for free/reduced price lunch. Although reliability for the WASL writing is relatively low (.72) it is difficult to understand how a teacher could judge a student who met the rigorous standard of the state's writing assessment as not on track to exit fifth grade.

Gay (1990) posits that even when teachers try conscientiously to control their ethnic, racial and social biases, they still may discriminate against culturally different students. The clustering of significant error rates for nonwhite males eligible for free and reduced priced lunch provided additional evidence that these biases exist and are operational in schools. (Mehan,Villanueva, Hubbard, and Lintz, 1996.) One explanation suggests the lack of success of poor minority males is related to their lack of "cultural capital". This "cultural capital" consists of the knowledge and familiarity with dominant uses of language, types of writing, and cultural and literary allusion which are transmitted through the family. Gaining and maintaining access to and mastery of the curriculum is dependent upon the students' possession of "cultural capital" and lack of it limits their chances to learn from educational material and interact profitably with teachers. In this study, perhaps teachers are influenced by this lack when they look at student work, failing to see the skills as presented.

The high Type 2 error rates for females, poor and not poor (.11 and .16 respectively) on WASL reading also provides evidence that teachers are responding to something other than students' actual work in making their judgments of proficiency. The WASL reading assessment invites some multiple choice items but includes two extensive constructed response tasks that require analysis and synthesis of information. It is highly unlikely that students could overscore on this assessment as statewide only 47.6 percent met the standard in reading. If girls can read, but their teachers indicate they cannot read, it points to the phenomenon studied by Sadker (1994), Oakes (1990) and others. Harvey (1986) and Sadker and Sadker (1986) reported that minority females receive the least attention in the classroom and most teachers are not aware of their own inequitable interactions with females. Receiving less attention, females are less likely than males to have opportunities to respond to openended questions and exhibit high order thinking skills. Harvey and the Sadkers found that brief, focused teacher training can reduce or eliminate these inequities -- which underscores the unintentionality of this inequity. This training is essential because if teachers do not expect that students can take part in a higher level discussion, those students are not even given a chance to participate (Stallings and McCarthy, 1990).

So the existence of low expectations (Brophy and Good, 1994) and lack of "cultural capital (Mehan, Villanueva, Hubbard and Lintz, 1996) inhibits the students' opportunities to learn. (Tettegah, 1997) refers to "cultural mismatch" between teachers and students, based on ethnic differences. The incidence of Type 2 error for poor females, regardless of race extends this discriminatory judgment to a group historically underserved by school.

If these students, poor white boys, poor minority boys and females are not judged proficient in classroom work when district and state evidence, particularly rigorous performance assessment indicates they have skills, , then one has to ask what are they doing in their classrooms on a daily basis? Does their classroom experience require them to show what they know in complex ways or are these students limited in their opportunities to respond? Are they in classrooms where they experience low level skill application?

Differential distribution of students to ability groups and tracks has been treated comprehensively by Oakes, Gamoran and Page (1992). The distribution of students to high, middle, and low ability groups or academic and general tracks seems to be related to ethnicity and socio economic status.

When aggregated at the school level, error types clustered by school. At a school level, 22 of 68 schools had a significant tendency to hold poor minority students and/or females behind who possibly, by virtue of test scores, have the necessary skills to be successful at the middle school level.

The nineteen schools that had a propensity to pass on students with low skill levels are indicative of the low expectations reported as school culture issues in the effective school literature. (Brookover, 1985).

In schools where the error rates are significant, it seems that teachers are looking at something beside skills, as exhibited in student work. They just don't know these students, don't know what they are doing and thus what they are capable of doing.

This is the first year of implementation and it is important to acknowledge that clear exemplars of student work are still not widely used as benchmarks in scoring student work and greater diffusion of models of good classroom based evidence is critical to the future success of the exit profile system. Additional clarity in grade level standards and proficiency is also required.

Conclusions

Equity issues in standard setting lie not so much in the standards but in the implementation and application of the standards. Their clarity provides the opportunity for equitable educational advancement regardless of race, gender, and social class only if all decision-makers can accurately judge those who reach standards and appropriately assist those who don't.

Educational equity should be conceptually understood as the comparability of learning opportunities and experiences to make high status knowledge and school success more accessible to students who are diversified by culture, ethnicity, class and gender. (Gay, 1990 p. 227)

It must be noted that in this district, teachers involved in making decisions about student learning have received minimum training and support in moving toward a standards based system. Over the last several years, political and economic forces have caused many teachers to retreat to safety behind their classroom door, while wars of resource allocation have let to sporadic unfocused staff development efforts. Only in direct writing assessment has there been any important effort in linking standards based teaching to scoring student work. New stable leadership from a visionary superintendent has provided the impetus for making this linkage.

Sarason writes "it is far easier to deal with villains than with well-intentioned educators imprisoned in tradition and by orientations that render self-scrutiny extraordinarily difficult." He acknowledges and it applies well to this large urban district, that teaching is a "taxing, frustrating, satisfying mind bending, and mind altering role for those who have not fallen prey to apathy and routine". (In Bartoli 1995, ix).

Due to an overreliance in the past with external norm referenced tests, it is likely that teachers in this district have not had reason to develop skills or confidence in their skills in judging student work.

Assessment measures must be developed that can illuminate the special talents of students of different ethnic, cultural, and linguistic backgrounds. (Lomax, et al 1996). Teachers must use these measures to increase expectations for underserved groups. In addition, thoroughly communicated district-wide grade level standards are necessary to provide accountability and clarity of expectation.

If as Gay (1990 p. 227) maintains, inequities are transmitted through irrelevant test content, testing structures and styles, teacher attitudes, instructional quality and program concentrations, then each and all of these transmission points must be addressed. It is this cluster of transmissions that creates an "ecology of inequity". The reversal of this ecology can only come about through extensive staff development, not through "narrow-minded accountability measures that encourage blame placing and denial of individual responsibility. (Bartoli, p. 139).

We must expand teachers'capacity to see skills embedded in student work and collect classroom based evidence that reflects district and state standards. We must also encourage schools to examine multiple forms of data and multiple representations of student work. No one piece of work or one test score can be a determinant of student progression in grade. (Carter, 1952).

We must also expand teachers strategies for providing opportunities to learn to all students regardless of race, class and gender. Understanding that the application of bias is unconscious, we should provide structured staff development in alternative teaching strategies such as cooperative learning, role playing, tutoring, team learning, demonstrating, coaching, problem solving and non directive teaching. Staff development must also be targeted toward helping teachers understand better how culturally different students go about the process of learning and demonstrating what they have learned. Teaching teachers to design, evaluate and use alternative evaluation techniques should also increase accuracy in judging student work. (Gay, 1990).

Continuing evaluation of the exit profile system is necessary to ensure equitable implementation and reduce disproportionality.

"Unless the teacher starts with a clear and realistic understanding of what students are and where they are coming from - what I have called the big but simple idea - they are doomed to feel inadequate and impotent, to frequently explaining their plight in the spirit of the dynamics of blaming the victim." (Sarason, in Bartoli, 1990, ix)

Ultimately, any policy construction regarding large scale application of standards in a high stakes system requires a definition of standards which included "opportunity for success" (Phillips, 1996). This requires that standards be a guarantee of standardized conditions that ensure that no students receive an unfair advantage or penalty rather that a guarantee of equal outcomes. As Stevens (1997) indicates, teacher bias and incompetency can distort the provision of standardized conditions and thus undermine any type of assessment.

We must focus on a renewal of instructional strategies and assessment, examination of personal assumption and biases, and reconnecting to families and communities. High standards can provide that focus. It is up to us to make sure they are high standards for everyone.

References

Bartoli, J. (1995). Unequal opportunity. (1995). New York: Teachers College Press.

Bronfenbrenner, U. (1988). Foreword. In A. Pence (Ed.) Ecological Research with Children and Families. New York.: Teachers College Press.

Brookover, W.B. (1985) Can we make schools effective for minority students?" The Journal of Negro Education, 54(3), 257-268.

Brophy, J. and Good T. (1994). Looking in Classrooms. New York: Harper Collins

Carter, R. (1952). How Invalid are Marks Assigned by Teachers? Journal of Educational Psychology, (43), 218-228.

Elmore, Richard and Associates (1991). Restructuring Schools: The Next Generation of School Reform. San Francisco: Jossey-Bass.

Gay, Geneva. (1990). Teacher preparation for equity. In Baptiste, H.P., Waxman, H.C., Wolkende, Felix J. and Anderson, J. (Eds.), Leadership Equity, School Effectiveness. Newbury Park, CA: Sage.

Goldenberg, C. & Gallemore, R. (1991). Local knowledge, research knowledge and educational change: A case study of early Spanish reading improvement." Educational Researcher, 20(8), 2-14.

Harvey G. (1986). Finding reality among the myths: Why what you thought about sex equity in education isn't so. Phi Delta Kappan, 67(7) 509-512.

Illinois State Board of Education. (1995). Performance Assessment in Mathematics: Approaches to Open-ended Problems. Springfield, IL: Author.

Koretz, D.M., Linn, R.L., Dunbar, S.B., and Shepard, L.A. (1991, April). The effects of high stakes testing on achievement: Preliminary findings about generalization Aaross Tests. Paper presented at the Annual meeting of the American Education Research Association, Chicago.

Linn R. (1994, October). The likely impact of performance standards as a function of uses: From rhetoric to sanctions. Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, Washington D.C.

Lomax, Richard G; West, M.M, Harmon, M.C., Viator , K.A. and Madaus, G. (1995). The impact of mandated standardized testing on minority students. Journal of Negro Education,64(2).

Madaus, G. (1992). A national testing system: Manna from above? Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Mehan, H., Villanueva, I., Hubbard, L. and Lintz, A., (1996). Constructing School Success: The Consequences of Untracking Low Achieving Students. London: Cambridge University Press.

Nicholls, J. & Hinkildsen, T. (Eds.) (1995). Reasons for Learning. New York: Teachers College Press.

Oakes, J. and Gamoran, A., and Page, R. (1992). Curriculum differentiation: Opportunities, outcomes and meanings. In Phillip Jackson (Ed.) Handbook of Research in Curriculum. pp. 570-608. New York: MacMillian

Oakes, Jeannie. (1990). Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science. Santa Monica, CA. Rand Corp.

Phillips, S. (1996). Legal defensibility of standards: Issues and policy perspectives." Educational Measurement,15 (2), 5-13.

Pinnell, G.S., Pikulski, J., Wixson, D.D., Campbell, J.R., Gough, P.B., and Beatty, S.S. (1995). Listening to Children Read Aloud. Washington, D.D.: OERI.

Sadker, M. and Sadker, D. (1986). Sexism in the classroom: From grade school to graduate school. Phi Delta Kappan 67(7), 512-515.

Sadker, Myra. (1994). Failing at fairness: How America's schools cheat girls. New York: C. Scribner's Sons.

Spandel, V. (1994). Seeing With New Eyes: A Guidebook on Teaching and Assessing Beginning Writers. Portland, OR: NWREL

Stallings, J. and McCarthy J. (1990). Instruction that enhances equity. In Baptiste (Ed.) School Effectiveness.

Stevens, F. (1997, February). Fairness in performance assessment: The impact of opportunity to learn on standards assessment outcomes. Paper presented at AASA Annual Meeting, Seattle, WA.

Stiggins, R. J. (1997). Student-centered Classroom Assessment. (2nd ed.). Upper Saddle River, NJ: Prentice-Hall

Tettegah, S. (1997). The racial consciousness attitudes of white prospective teachers and their perceptions of the teachability of students from different racial/ethnic backgrounds: Findings from a California study. Journal of Negro Education.

Washington Commission on Student Learning. (1997). Primary classroom-based assessment tool kit. Olympia, WA. Author.

Weinstein, R. (1996). Higher standards in a tracked system of schooling. Educational Researcher,26 (8) 16-18.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028358

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Addressing Validity ISSues in Student Exit Performance ASSESSMENT

Author(s): Jill Hearne and Madelaine Ramey

Corporate Source: Seattle Public Schools 815 4th Ave. N. Seattle WA. 98109

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Informetion Center (ERIC) nonexclusive permission to reproduce end disseminate this document es indicated ebove. Reproduction from the ERIC microfiche or electronic medie by persons other then ERIC employees end its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy informetion needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:

Printed Name/Position/Title: Dr. Jill Hearne Coordinator of Assessment

Organization/Address: Seattle Public Schools

Telephone: 206 298 7235

FAX: 206 298 7131

E-Mail Address: jhearne@cks.ssd.K12.WA.US

Date: 4/7/98

(over)

ERIC Full Text Provided by ERIC

ERIC® **Clearinghouse on Assessment and Evaluation**

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

March 20, 1998

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae.net.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (424)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1998/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an AERA chair or discussant, please save this form for future use.

CUA

The Catholic University of America

ERIC Full Text Provided by ERIC